

Quantitative Group Testing for Heavy Hitter Detection

Chao Wang, Qing Zhao, Chen-Nee Chuah
Department of Electrical and Computer Engineering,
University of California, Davis, CA 95616
{eecwang, qzhao, chuah}@ucdavis.edu

Abstract— We consider the quantitative group testing problem where the objective is to identify defective items in a given population based on results of tests performed on subsets of the population. Under the quantitative group testing model, the result of each test reveals the number of defective items in the tested group. The minimum number of tests achievable by nested test plans was established by Aigner and Schughart in 1985 within a minimax framework. The optimal nested test plan offering this performance, however, was not obtained. In this work, we establish the optimal nested test plan in closed form. This optimal nested test plan is also asymptotically (as the population size grows to infinity) optimal among all test plans. We then focus on the application of heavy hitter detection problem for traffic monitoring and anomaly detection in the Internet and other communication networks. For such applications, it is often the case that a few abnormal traffic flows with exceptionally high volume (referred to as heavy hitters) make up most of the traffic seen by the entire network. Since the volume of heavy hitters is much higher than that of normal flows, the number of heavy hitters in a group of flows can be accurately estimated from the aggregated traffic load. Other potential applications include detecting idle channels in the radio spectrum in the high SNR regime.

Index Terms—Group testing, quantitative group testing, heavy hitter detection, anomaly detection, traffic measurements, spectrum sensing.

I. INTRODUCTION

A. Group Testing

The group testing problem is concerned with identifying defective items in a given population by performing tests over subsets of the population. Under the classic model, each test gives a binary result, indicating whether the tested group contains any defective items. The objective is a test plan that minimizes the number of tests required for identifying all defective items.

The problem was first motivated by the practice of screening draftees with syphilis during World War II, and the idea of testing pooled blood samples from a group of people (rather than testing each person one by one) was initiated by Robert Dorfman [1]. In Dorfman's test plan, draftees are tested in groups with a suitable size. If a group is tested positive, its members are tested one by one to identify the infected individual(s). An improvement to Dorfman's test plan was

proposed by Sterrett in 1957 [2]. The improvement suggested that once an infected person from a group is identified, the rest of the group is again tested together.

General formulations of and rigorous attacks on group testing were pioneered by Sobel and Groll in their paper published in 1959 [3]. Sobel and Groll adopted a probabilistic model on the defective items and focused on the problem of minimizing the expected number of tests. This formulation of group testing was later known as *probabilistic group testing* (PGT). Recognizing the intractability of the optimal solution to the general problem, Sobel and Groll considered a class of test plans with a *nested* structure. Specifically, in a nested test plan, once a test reveals a defective group, the next test must be on a proper subset of this group. Sobel and Groll characterized implicitly the optimal nested test plan with a pair of recursive formulas and solved them numerically. They also established several asymptotic (as the population size approaches infinity) properties of the optimal nested test plan.

The counterpart to PGT is the *combinatorial group testing* (CGT) formulated and studied by Li [4] and Katona [5]. In CGT, there are n items among which d are defective. There is no probabilistic knowledge on the defective sets, and the objective is to minimize the number of tests in the worst case (i.e., a minimax formulation rather than a Bayesian formulation as in PGT) [6].

Under both formulations, the test plans can be adaptive or non-adaptive. Adaptive test plans are sequential in nature: which group to test next depends on the outcome of the previous tests. The studies in [3]–[5] mentioned above all focus on adaptive test plans. Non-adaptive group testing is a one-stage problem in which all actions can be determined before any test is performed. Non-adaptive test plans are often represented by matrices [7], [8].

The classic group testing formulation has seen a wide range of applications, including chemical apparatus leakage detection [3], multiaccess communications [9]–[11], idle channel detection in the radio spectrum [12], compressed sensing [13], network tomography [14] and anomaly detection [15], [16]. In particular, non-adaptive group testing has been widely applied to DNA sequencing and DNA library screening [7], [17]–[19].

B. Quantitative Group Testing for Heavy Hitter Detection

In this work, we consider quantitative group testing in which a test reveals not only whether the tested group is contami-

nated, but also the number of defective items in the tested group. Under the combinatorial formulation of the problem, Aigner and Schugart established in [20] the performance (i.e., the number of required tests) of the optimal nested test plan. The optimal nested test plan itself, however, was not obtained. In this work, we establish the optimal nested test plan in closed form. This optimal nested test plan is also asymptotically (as the population size grows to infinity) optimal among all test plans.

We then focus on the application of quantitative group testing to the heavy hitter detection problem for traffic monitoring and anomaly detection in the Internet and other communication networks. For Internet traffic, it is a common observation that a small percentage of high-volume flows (referred to as heavy hitters) account for most of the total traffic [21]. In particular, it was shown in [22] that the top (in terms of volume) 9% of flows make up 90.7% of the total traffic over the Internet. Quickly identifying the heavy hitters is thus crucial to network stability and security. However, the large number of Internet flows makes individual monitoring extremely inefficient if not impossible. The quantitative group testing model stems from the fact that the difference between the average traffic rates of heavy hitters and normal flows is large, which allows for accurate estimation of the number of heavy hitters from the aggregated traffic load. Through simulation examples, we examine the impact of the estimation error (caused by the random nature of the traffic volume of both normal flows and heavy hitters) in each group test on the end result of heavy hitter detection in terms of false positive and false negative rates. Other potential applications include detecting idle channels in the radio spectrum when the signal strength is relatively even across busy channels and much higher than the noise level in idle channels (the high SNR regime).

C. Related Work

Quantitative group testing is one of the several models under the so-called additive group testing problems [6]. It is also known as the coin weighing problem with a spring scale first introduced by Shapiro in 1960 [23]. The problem is to identify d counterfeit coins in a collection of n . The weights of the authentic and counterfeit coins are known. Thus each weighing gives the number of counterfeit coins in the tested group. Most studies on this problem focus on non-adaptive test plans, see, for example, [24]–[32] on the case of unknown d and [29], [33], [34] on the case of known d . On adaptive test plans for quantitative group testing, there are a number of results on the special case of $d = 2$ (see [28], [29], [35]–[38]). For the general case with $0 < d < n$, besides the work by Aigner and Schugart [20] discussed in Section I-B, Bshouty developed a polynomial-time algorithm with a performance no worse than twice of the information-theoretic lower bound [39].

The applications of quantitative group testing include the uniquely decodable codes for noiseless n -user adder channel problem [40], and the construction of unknown graphs from

additive queries [34], [36], [41]. Several variations of the problem can be found in [42]–[44].

On the heavy hitter detection problem, most existing work falls into the category of streaming algorithms that sample the incoming stream of packets to maintain a counter for heavy hitter suspects that are updated dynamically (see the counter-based algorithms given in [45]–[47]) or a sketch of all flows based on frequency estimation (see the sketch-based algorithms given in [48]–[50]). Other types of heavy hitter detection schemes include per-flow based traffic monitoring [51], [52] and application-oriented approaches [53], [54]. To our best knowledge, detecting heavy hitters by measuring aggregated traffic under a quantitative group testing formulation is new.

II. PROBLEM FORMULATION

Under the CGT formulation, we are given a population of n items, each labeled with a unique ID. It is known that among these n items, d are defective. We assume that $1 \leq d \leq n - 1$ to avoid the trivial scenarios of $d = 0$ and $d = n$ and use (n, d) to denote a specific CGT problem.

For a given test plan π , the number of tests required by π to identify all d defective items in a population of size n depends on which d items are defective. Let $N_\pi(n, d; \mathcal{D})$ denote the number of tests required by π when the d defective items are given in the set \mathcal{D} . Note that n and d are known while \mathcal{D} is unknown and is what the test plan needs to identify. Under the combinatorial formulation, the performance of a test plan is determined by the worst instant of \mathcal{D} among all subsets with size d . The performance of π , denoted by $N_\pi(n, d)$, is thus given by

$$N_\pi(n, d) = \max_{\mathcal{D} \subset (n), |\mathcal{D}|=d} N_\pi(n, d; \mathcal{D}), \quad (1)$$

where (n) denote the entire population. Our objective is an optimal nested test plan π^* given by

$$\pi^* = \arg \min_{\pi \in \Pi} N_\pi(n, d), \quad (2)$$

where Π denotes the family of all admissible nested test plans. To simplify the notation, the performance of the optimal nested test plan π^* is denoted by $N(n, d)$ (rather than $N_{\pi^*}(n, d)$), which will also be referred to as the optimal number of tests for identifying d defective items in the population. Let $M(n, d)$ denote the optimal size of the first group test for (n, d) . The value of $M(n, d)$ for all n and d fully specifies the optimal nested test plan π^* .

Considering the minimax nature of the CGT formulation, we arrive at the following recursive formula for the optimal number of tests:

$$N(n, d) = \min_{m=1, \dots, n} \left\{ \max_{d_1 = \max\{0, d+m-n\}, \dots, \min\{m, d\}} \{1 + N(m, d_1) + N(n-m, d-d_1)\} \right\}. \quad (3)$$

Define

$$M(n, d) = \arg \min_{m=1, \dots, n} \left\{ \max_{d_1 = \max\{0, d+m-n\}, \dots, \min\{m, d\}} \{1 + N(m, d_1) + N(n-m, d-d_1)\} \right\} \quad (4)$$

The recursions of $N(n, d)$ and $M(n, d)$ were once considered by Aigner and Schughart in [20], where they gave the closed form of $N(n, d)$ for all n and d . However the test plan, $M(n, d)$, was not explicit for all n and d . In this work, by solving the integer optimization problem defined in (3) and (4), we show that the CGT in quantitative model admits a clean solution in closed form as given in the following section.

Fig. 1. The logarithmic order with n of $N(n, d)$ ($d = 5, n \geq 10$).

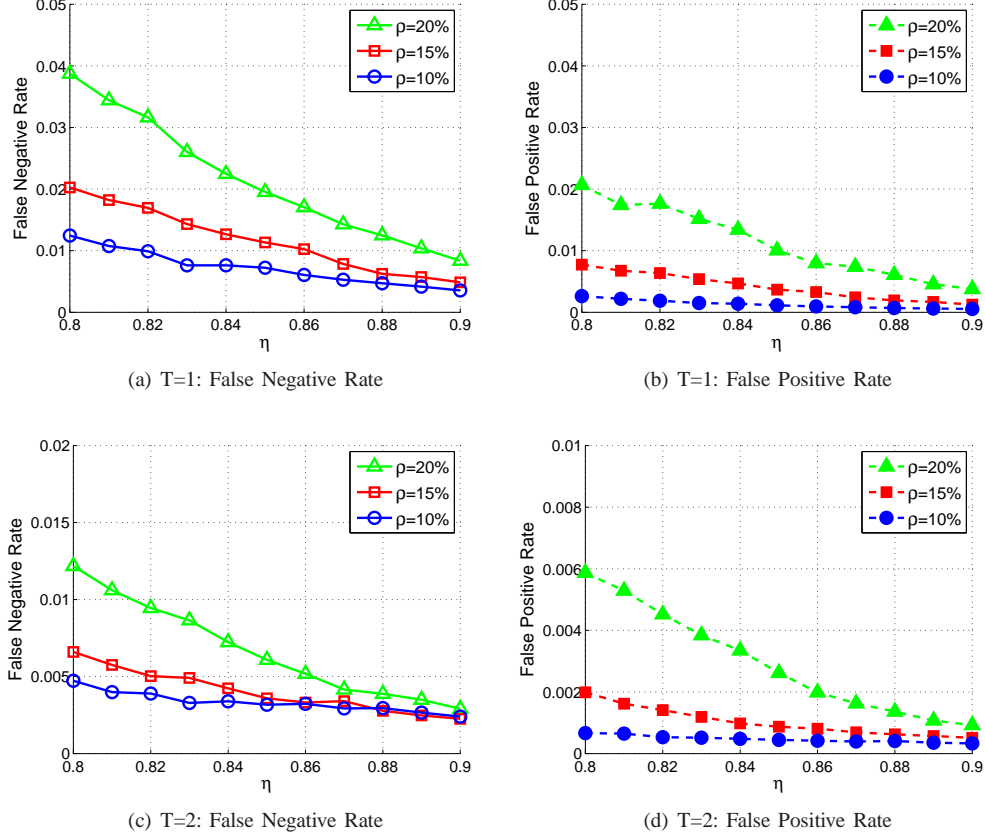


Fig. 2. False Negative and False Positive Rates versus η - Poisson Distribution with ML Estimator

Next, we show the order of the optimal number of tests in terms of n and d . We can rewrite (5) as follows:

$$N(n, d) = \underbrace{\lceil \log_2 \frac{n}{d} \rceil \cdot d}_{P_1} + \underbrace{\lceil \frac{n}{2^l} \rceil - d - 1}_{P_2}, \quad (9)$$

where l is given in (7). It is easy to see that the second term P_2 is bounded in n . We subsequently conclude that the optimal nested test plan guarantees to identify all d defective items in a population of n in $O(d \log_2(n/d))$ tests, which is logarithmic with n . Fig. 1 illustrates the scaling behavior of $N(n, d)$ in n . By comparing with the information theoretic lower bound, this query complexity is asymptotically optimal over all algorithms when d is a constant and the population size grows to infinity.

IV. CGT WITHOUT PRIOR KNOWLEDGE

We have so far focused on the standard CGT formulation which assumes a prior knowledge on the total number of defective items in the given population. For applications where this prior knowledge is unavailable, the first question is how to start the first test: for any population size n , should the first test be carried over the entire population or a proper subset of the population with the size potentially depending on n ? In the theorem below, we show that within the class of nested test plan, the optimal action in the first step is to test the entire population. The first test will then reveal the total number d

of defective items, and the problem is reduced to a CGT of (n, d) .

Theorem 2: For a given population with any size n , within the class of nested test plans, the optimal action in the first step is to test the entire population.

V. APPLICATION TO HEAVY HITTER DETECTION

In this section, we consider the application of the optimal nested test plan developed in Section III to the heavy hitter detection problem. Consider a network consisting of n flows, among which d are heavy hitters. We are going to study two different probabilistic models of flows. First, we assume that each flow is an independent Poisson distribution with rate μ_0 for normal flows and μ_1 for heavy hitters. Define

$$\rho = \frac{d}{n}, \quad \eta = \frac{d\mu_1}{d\mu_1 + (n-d)\mu_0} \quad (10)$$

as the fraction of heavy hitters in terms of number of flows and the total traffic volume, respectively. Over the Internet, we typically have ρ around 10% to 20% and η around 80% to 90%.

To apply the group testing plan to the heavy hitter detection problem, we need to estimate the number of heavy hitters from measurements of aggregated traffic rate. Assume that m flows are aggregated together and T measurements are taken. Based on the independent Poisson assumption on each

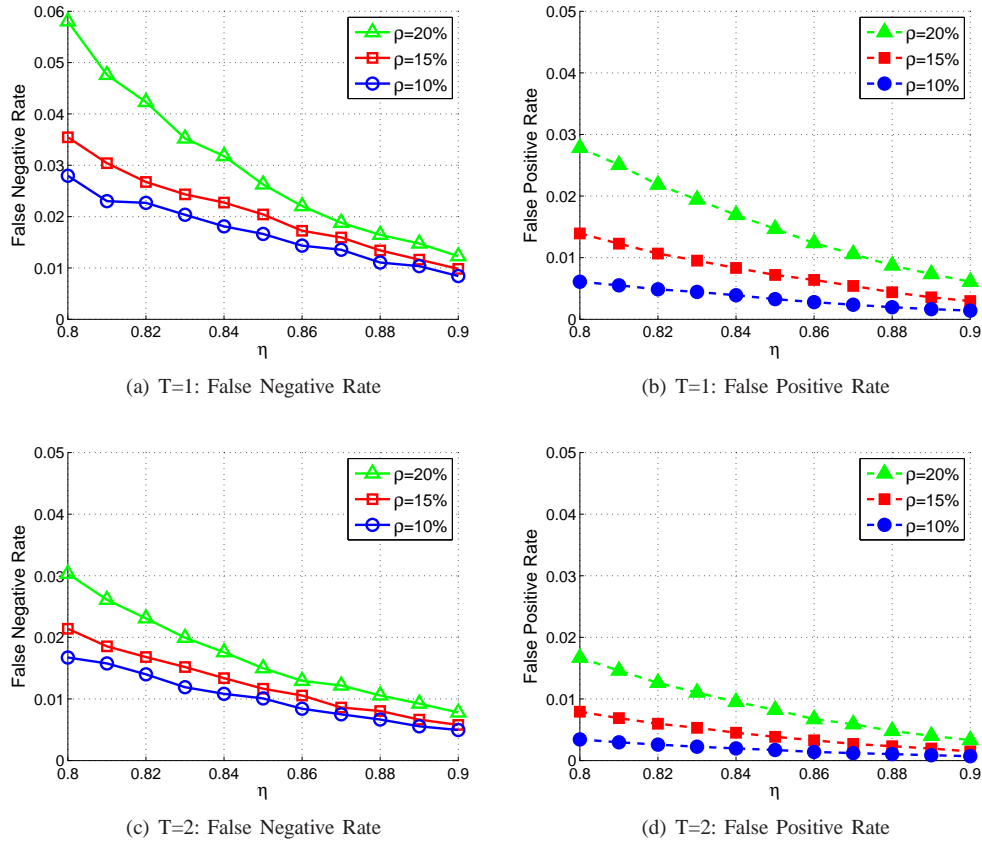


Fig. 3. False Negative and False Positive Rates versus η - Log-normal Distribution with Sample Mean Estimator

flow, the aggregated traffic is again Poisson distributed with rate $d_1\mu_1 + (m - d_1)\mu_0$. From the T measurements of the aggregated traffic, we can obtain a Maximum Likelihood (ML) estimate of the number d_1 of heavy hitters among this group of m flows, which can be written as

$$\hat{d}_1 = \arg \max_{0 \leq d_1 \leq m} \frac{\sum_{i=1}^T z_i}{T} [\log(d_1\mu_1 + (m - d_1)\mu_0)] - [d_1\mu_1 + (m - d_1)\mu_0], \quad (11)$$

where z_i is observation of traffic rate in i th measurement. This estimated value \hat{d}_1 will be used as the outcome of this group test, which will determine the size of the next group test based on the optimal nested test plan given in Theorem 1.

In the second numerical example, we assume that each flow is an independent log-normal distribution with mean value μ_0 for normal flows and μ_1 for heavy hitters. The definitions in (10) are still valid. But different to the Poisson model, since the likelihood function is not in closed form, now we are using sample mean estimator to estimate the number of heavy hitters in the aggregated group. i.e.

$$\hat{d}_1 = \left\lceil \frac{\sum_{i=1}^T z_i / T - m\mu_0}{\mu_1 - \mu_0} \right\rceil, \quad (12)$$

where “[.]” denotes the operation of taking the nearest integer.

Fig.2 and Fig.3 show the simulation examples on the two types of detection errors for various values of ρ and η . For

convenience, we keep $\mu_0 = 1$, so that the value of μ_1 will change with ρ and η according to (10). It is obvious that, for fixed ρ , when η increases, the false negative and false positive rates will both decrease. When η is fixed, the groups with smaller ρ will have lower detection errors. It’s easy to see from (10) that, when ρ decreases and η increases, the difference between μ_1 and μ_0 will become larger. It makes the mean square error of the ML estimator and sample mean estimator smaller. That’s why the estimation of d_1 become more accurate and probabilities of detection error will decrease. On the other hand, by comparing (a)(b) with (c)(d) in these figures, we can see when the value of T in each test changes from 1 to 2, the estimation error of d_1 becomes smaller with more measurements, therefore the final detection errors will decrease significantly.

VI. CONCLUSIONS

We studied the quantitative group testing problem within the combinatorial group testing framework. The optimal nested test plan is established in closed form. The result finds applications in heavy hitter detection and spectrum sensing.

REFERENCES

- [1] R. Dorfman, “The detection of defective members of large populations,” *The Annals of Mathematical Statistics*, vol. 14, no. 4, pp. 436–440, 1943. [Online]. Available: <http://www.jstor.org/stable/2235930>

- [2] A. Sterrett, "On the detection of defective members of large populations," *The Annals of Mathematical Statistics*, vol. 28, no. 4, pp. 1033–1036, 1957. [Online]. Available: <http://www.jstor.org/stable/2237067>
- [3] M. Sobel and P. A. Groll, "Group testing to eliminate efficiently all defectives in a binomial sample," *Bell System Technical Journal*, vol. 38, no. 5, pp. 1179–1252, 1959.
- [4] C. H. Li, "A sequential method for screening experimental variables," *Journal of the American Statistical Association*, vol. 57, no. 298, pp. 455–477, 1962.
- [5] G. O. Katona, "Combinatorial search problems," *A survey of combinatorial theory*, pp. 285–308, 1973.
- [6] D. Du and F. Hwang, *Combinatorial group testing and its applications*, 2nd ed. World Scientific, 2000.
- [7] H. Q. Ngo and D.-Z. Du, "A survey on combinatorial group testing algorithms with applications to DNA library screening," *Discrete mathematical problems with medical applications*, vol. 55, pp. 171–182, 2000.
- [8] D. Du and F. Hwang, *Pooling Design and Nonadaptive Group Testing: Important Tools for DNA Sequencing*. World Scientific, 2006.
- [9] J. Wolf, "Born again group testing: Multiaccess communications," *IEEE Transactions on Information Theory*, vol. 31, no. 2, pp. 185–191, Mar 1985.
- [10] J. K. Wolf, "Principles of group testing and an application to the design and analysis of multi-access protocols," in *The Impact of Processing Techniques on Communications*. Springer, 1985, pp. 237–257.
- [11] T. Berger, N. Mehravari, D. Towsley, and J. Wolf, "Random multiple-access communication and group testing," *IEEE Transactions on Communications*, vol. 32, no. 7, pp. 769–779, Jul 1984.
- [12] A. Sharma and C. Murthy, "Group testing based spectrum hole search for cognitive radios," *IEEE Transactions on Vehicular Technology*, vol. PP, no. 99, pp. 1–1, 2014.
- [13] M. Cheraghchi, A. Hormati, A. Karbasi, and M. Vetterli, "Group testing with probabilistic tests: Theory, design and application," *IEEE Transactions on Information Theory*, vol. 57, no. 10, pp. 7057–7067, Oct 2011.
- [14] M. Cheraghchi, A. Karbasi, S. Mohajer, and V. Saligrama, "Graph-constrained group testing," *IEEE Transactions on Information Theory*, vol. 58, no. 1, pp. 248–262, Jan 2012.
- [15] M. T. Thai, Y. Xuan, I. Shin, and T. Znati, "On detection of malicious users using group testing techniques," in *The 28th International Conference on Distributed Computing Systems*. IEEE, 2008, pp. 206–213.
- [16] S. Khatlab, S. Gobriel, R. Melhem, and D. Mosse, "Live baiting for service-level DoS attackers," in *The 27th Conference on Computer Communications*. IEEE, April 2008.
- [17] W. J. Bruno, E. Knill, D. J. Balding, D. Bruce, N. Doggett, W. Sawhill, R. Stallings, C. C. Whittaker, and D. C. Torney, "Efficient pooling designs for library screening," *Genomics*, vol. 26, no. 1, pp. 21–30, 1995.
- [18] D. Balding, W. Bruno, D. Torney, and E. Knill, "A comparative survey of non-adaptive pooling designs," in *Genetic mapping and DNA sequencing*. Springer, 1996, pp. 133–154.
- [19] M. Farach, S. Kannan, E. Knill, and S. Muthukrishnan, "Group testing problems with sequences in experimental molecular biology," in *Proceedings of Compression and Complexity of Sequences*. IEEE, 1997, pp. 357–367.
- [20] M. Aigner and M. Schugart, "Determining defectives in a linear order," *Journal of Statistical Planning and Inference*, vol. 12, pp. 359–368, 1985.
- [21] K. Thompson, G. Miller, and R. Wilder, "Wide-area internet traffic patterns and characteristics," *IEEE Network*, vol. 11, no. 6, pp. 10–23, Nov 1997.
- [22] W. Fang and L. Peterson, "Inter-AS traffic patterns and their implications," in *Global Telecommunications Conference*, vol. 3, 1999, pp. 1859–1868.
- [23] H. S. Shapiro, "Problem E 1399," *Amer. Math. Monthly*, vol. 67, no. 82, pp. 697–697, 1960.
- [24] N. Fine, "Solution of problem E 1399," *American Mathematical Monthly*, vol. 67, no. 7, pp. 697–698, 1960.
- [25] D. G. Cantor, "Determining a set from the cardinalities of its intersections with other sets," *Canadian Journal of Mathematics*, vol. 16, pp. 94–97, 1964.
- [26] B. Lindström, "On a combinatorial problem in number theory," *Canad. Math. Bull.*, vol. 8, no. 4, pp. 477–490, 1965.
- [27] —, "On a combinatorial detection problem II," *Studia Scientiarum Mathematicarum Hungarica*, vol. 1, pp. 353–361, 1966.
- [28] —, "On möbius functions and a problem in combinatorial number theory," *Canad. Math. Bull.*, vol. 14, no. 4, pp. 513–516, 1971.
- [29] B. Lindström et al., "Determining subsets by unramified experiments," 1975.
- [30] P. Erdős and A. Rényi, "On two problems of information theory," 1963.
- [31] S. Soderberg and H. S. Shapiro, "A combinatorial detection problem," *American Mathematical Monthly*, pp. 1066–1070, 1963.
- [32] D. G. Cantor and W. Mills, "Determination of a subset from certain combinatorial properties," *Canadian Journal of Mathematics*, vol. 18, pp. 42–48, 1966.
- [33] A. Djakov, "On a search model of false coins," in *Topics in Information Theory (Colloquia Mathematica Societatis Janos Bolyai 16, Keszthely, Hungary)*. Budapest, Hungary: Hungarian Acad. Sci., 1975, p. 163170.
- [34] S.-S. Choi and J. H. Kim, "Optimal query complexity bounds for finding graphs," in *Proceedings of the 40th annual ACM symposium on Theory of computing*. ACM, 2008, pp. 749–758.
- [35] C. Christen, "A fibonacci algorithm for the detection of two elements," vol. Publ. 341, Dept. d'IRO, 1980.
- [36] M. Aigner, "Search problems on graphs," *Discrete Applied Mathematics*, vol. 14, no. 3, pp. 215–230, 1986.
- [37] F. H. Hao, "The optimal procedures for quantitative group testing," *Discrete Applied Mathematics*, vol. 26, no. 1, pp. 79–86, 1990.
- [38] L. Gargano, V. Montouri, G. Setaro, and U. Vaccaro, "An improved algorithm for quantitative group testing," *Discrete applied mathematics*, vol. 36, no. 3, pp. 299–306, 1992.
- [39] N. H. Bshouty, "Optimal algorithms for the coin weighing problem with a spring scale," in *COLT*, 2009.
- [40] S.-C. Chang and E. Weldon, "Coding for T-user multiple-access channels," *IEEE Transactions on Information Theory*, vol. 25, no. 6, pp. 684–691, 1979.
- [41] V. Grebinski and G. Kucharov, "Optimal reconstruction of graphs under the additive model," *Algorithmica*, vol. 28, no. 1, pp. 104–124, 2000.
- [42] N. H. Bshouty and H. Mazzawi, "Toward a deterministic polynomial time algorithm with optimal additive query complexity," in *Mathematical Foundations of Computer Science*. Springer, 2010, pp. 221–232.
- [43] V. Grebinski and G. Kucharov, "Optimal reconstruction of graphs under the additive model," in *Algorithms-ESA'97*. Springer, 1997, pp. 246–258.
- [44] W. Han, P. I. Frazier, and B. M. Jedynak, "Twenty questions for localizing multiple objects by counting: Bayes optimal policies for entropy loss," *arXiv preprint arXiv:1407.4446*, 2014.
- [45] G. S. Manku and R. Motwani, "Approximate frequency counts over data streams," in *Proceedings of the 28th international conference on Very Large Data Bases*. VLDB Endowment, 2002, pp. 346–357.
- [46] A. Metwally, D. Agrawal, and A. El Abbadi, "Efficient computation of frequent and top-k elements in data streams," in *Database Theory-ICDT*. Springer, 2005, pp. 398–412.
- [47] E. D. Demaine, A. López-Ortiz, and J. I. Munro, "Frequency estimation of internet packet streams with limited space," in *Algorithms ESA*. Springer, 2002, pp. 348–360.
- [48] G. Cormode and S. Muthukrishnan, "An improved data stream summary: the count-min sketch and its applications," *Journal of Algorithms*, vol. 55, no. 1, pp. 58–75, 2005.
- [49] C. Estan and G. Varghese, *New directions in traffic measurement and accounting*. ACM, 2002, vol. 32, no. 4.
- [50] G. Cormode and S. Muthukrishnan, "What's hot and what's not: tracking most frequent items dynamically," *ACM Transactions on Database Systems (TODS)*, vol. 30, no. 1, pp. 249–278, 2005.
- [51] M. Al-Fares, S. Radhakrishnan, B. Raghavan, N. Huang, and A. Vahdat, "Hedera: Dynamic flow scheduling for data center networks," in *NSDI*, vol. 10, 2010, pp. 19–19.
- [52] N. Farrington, G. Porter, S. Radhakrishnan, H. H. Bazzaz, V. Subramanya, Y. Fainman, G. Papen, and A. Vahdat, "Helios: a hybrid electrical/optical switch architecture for modular data centers," *ACM SIGCOMM Computer Communication Review*, vol. 41, no. 4, pp. 339–350, 2011.
- [53] R. Braden, D. Clark, S. Shenker et al., "Integrated services in the internet architecture: an overview," 1994.
- [54] M. Roughan, S. Sen, O. Spatscheck, and N. Duffield, "Class-of-service mapping for QoS: a statistical signature-based approach to IP traffic classification," in *Proceedings of the 4th ACM SIGCOMM conference on Internet measurement*. ACM, 2004, pp. 135–148.